# The Partially-Matched-Sample Correction in Pseudo Panel Minimum Distance Estimation

Fei Jia*

July 24, 2024

**Abstract**

Certain repeated cross-sectional data sets, such as the Current Population Survey (CPS), use special sampling designs by which samples from different times periods are partially matched. This paper proposes a correction to the optimal weighting matrix in minimum distance (MD) estimation of pseudo panel models to account for such partially matched samples. This partially-matched-sample correction may be needed if the sample matching rate is nontrivial and, at the same time, there is a fixed effect, a serially correlated idiosyncratic error, or both in the underlying linear panel data model data generating process, all of which lead to a block diagonal structure of the optimal weighting matrix. Using the correction can result in considerable efficiency gains both in finite sample and asymptotically. As an illustration, the correction is applied to the classical question of estimating the monetary return to education using the yearly Merged Outgoing Rotation Group (MORG) files from CPS.

## 1 Introduction

Since the seminal work of Deaton (1985), the pseudo panel approach has become a widely used alternative to standard panel data methods for estimating an underlying linear panel data model with unobserved individual fixed effects when only repeated cross sections are available. The key idea of the approach is to group individuals into cohorts according to certain fixed characteristics,[1] such as year of birth, to construct a synthetic panel at the cohort level. Subsequently, these cohorts are tracked overtime as research units. The constructed panel is "pseudo" because measures regarding the cohorts are constructed from variable sample cohort means rather than directly observed. While often regarded as an alternative type of data sets, repeated cross sections offer several advantages over traditional panel data sets. These advantages

*Department of Economics, Richard A. Chaifetz School of Business, Saint Louis University. Email: fei.jia@slu.edu

[1] Groups and cohorts are used interchangeably in this paper.

include more abundant and accessible data, generally larger sample sizes, and natural immunity to attrition. Such strengths render pseudo panel models versatile and applicable to various research domains spanning economics, political science, epidemiology, etc.. In economics, for instance, they are particularly useful when studying labor supply (Browning et al. 1985, Blundell et al. 1993 and Campbell & Lusher 2019), economic mobility (Antman & McKenzie 2007, Dang et al. 2014 and Dang & Lanjouw 2023), commodity demand (Browning et al. 1985, Gardes et al. 2005 and Meng et al. 2014), health (Saksena & Maldonado 2017), returns to education (Moretti 2004 and Jones et al. 2023), and other topics where observing the same entities over time is challenging.

On the theoretical front, a variety of developments have also been made in the literature to broaden the scope of application scenarios for pseudo panel models. Notable extensions include dynamic models (Moffitt 1993, Collado 1997 and Verbeek & Vella 2005), alternative asymptotic dimensions other than large cohort size (Verbeek & Nijman 1993 and Collado 1997), unequally spaced pseudo panels (McKenzie 2001), heterogeneity of fixed effects within cohorts (McKenzie 2004), estimation framework and asymptotic efficiency (Imbens & Wooldridge 2007 and Inoue 2008), and cohort interactive effects (Juodis 2018). Some systematic discussions on the theoretical development of pseudo panel models can be found in Imbens & Wooldridge (2007), Inoue (2008) and Verbeek (2008), among others.

Despite the rich literature on pseudo panel models, existing papers tend to rely on a common sampling design assumption that the samples from different time periods are independent of each other, termed as serially independent sampling henceforth. When combined with random sampling within each cross section, this assumption implies that (the inverse of) the optimal weighting matrix in either minimum distance (MD) or generalized method of moments (GMM) estimation is diagonal. While serial independent sampling is appropriate for many cross-sectional data sets used in the literature, there are important exceptions where samples from different periods are partially matched due to specific survey designs.[2] For instance, the Current Population Survey (CPS) employs a special rotation group design (see CPS Technical Documentation 2014), resulting in partially matched samples over consecutive months or years. The derivative Merged Outgoing Rotation Group (MORG) files, which includes only the CPS data obtained at the 4th and 8th interviews for each rotation group, also contain partially matched samples over time. Likewise, the Australia Labour Force Survey (LFS) that rotates out 1/8 of the sample each month also consists of partially matched samples. In such repeated cross-sectional data sets with partially matched samples, the dependence among samples introduces potential serial correlation in the cohort-level composite errors of the pseudo panel model. Consequently, the standard diagonal weighting matrix under serially independent samples is no longer optimal,

[2] It is worth noting that partially matched samples may also arise from a finite population combined with a nontrivial sampling rate, even if samples are serially independent. Intuitively, this type of sample overlaps by chance does not require any additional treatment.

although estimators based on such standard weighting remain consistent.

This paper explores this feature of partially matched samples in repeated cross sections and proposes a correction to the standard weighting matrix in pseudo panel MD estimation to restore its optimality. The inverse of the proposed optimal weighting matrix is found to be block diagonal, sharing the same diagonal elements as the inverse of the standard weighting matrix but potentially featuring nonzero off-diagonal elements. An asymptotic theory tailored to such partially matched samples under large number of cohort observations, fixed number of cohorts and fixed number of time periods is developed. Regarding (asymptotic) efficiency, the gain from using this correction is substantial when partially matched samples lead to a significant deviation between the optimal and standard weightings. A series of meticulously designed simulation cases reveal potential features in the underlying data generating process (DGP) that may induce such differences. These features include the relative magnitude of the group-time cell variance of the fixed effect with respect to (w.r.t.) that of the idiosyncratic error, cohort-wise heteroskedasticity in the fixed effect, cell-wise heteroskedasticity, and/or serial correlation in the idiosyncratic error. To illustrate the proposed correction, the paper revisits the classical inquiry of estimating the monetary return to education using the MORG files from 2010 to 2019. The analysis demonstrates that the proposed optimal weighting that accounts for the partial sample matching yields approximately 10% smaller standard errors (s.e.'s hereafter) compared to the standard optimal weighting that ignore the matching.

The rest of the paper is organized as follows. Section 2 presents the MD framework for estimating pseudo panels and reviews the standard optimal MD weighting in the literature. In Section 3, the partially-matched-sample correction to the optimal weighing matrix is derived. Section 4 shows the simulation study. Section 5 illustrates the proposed correction by applying it to the classical empirical question of estimating the monetary return to education using the MORG files. Section 6 concludes.

## 2 Minimum Distance Estimation of Pseudo Panel Models

The MD framework utilized in this paper draws heavily from Imbens & Wooldridge (2007), albeit with certain modifications. Although it may seem distinct, the MD framework shares fundamental similarities with the GMM framework employed by Inoue (2008), as both rely on precisely the same set of conditional moment conditions and yield identical estimators and statistical inference. The decision to adopt the MD framework in this paper stems from its inherent compatibility with pseudo panel models.[3]

[3] As Imbens & Wooldridge (2007) points out, the core MD idea of extracting structural parameters from easily estimable reduced-form parameters aligns naturally with pseudo panel models. Additionally, the MD framework boasts several strengths, including its ability to separate the population model from sampling assumptions, elucidate why exogeneity of the group membership, rather than exogeneity of the regressors, is required for consistency, and, perhaps most importantly, reveal the key identification condition that there must be enough group-time variation in the group-time cell means of the covariates.

## 2.1 The population pseudo panel model

This subsection exclusively presents the population model. Throughout paper, the underlying individual-level data generating process (DGP) for individual $i$ over $T$ time periods is assumed to be the linear fixed-effects (FE) panel data model

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{d}_t\boldsymbol{\eta}' + f_i + u_{it},\ t = 1, \dots, T \tag{2.1}$$

where $y_{it}$ is a scalar dependent variable, $\mathbf{x}_{it}$ is a $1 \times K$ vector of regressors that contains unity on its first entry, $\mathbf{d}_t \equiv (d_{2t}, ..., d_{Tt})$ with $d_{st} = 1_{\{s=t\}}$ is a $1 \times (T-1)$ vector of time dummies,[4] $f_i$ is a scalar individual-specific fixed effect that is allowed to be correlated with $\mathbf{x}_{it}$, and $u_{it}$ is a scalar idiosyncratic error; $\boldsymbol{\beta} \equiv (\beta_1, \dots, \beta_K)'$ is the $K \times 1$ parameter vector for $\mathbf{x}_{it}$, and $\boldsymbol{\eta} \equiv (\eta_2, \dots, \eta_T)'$ is the $(T-1) \times 1$ parameter vector of time effects.[5] The dimensions of $\mathbf{d}_t$ and $\boldsymbol{\eta}$ are $T-1$ because $\mathbf{x}_{it}$ contains unity and, as a result, the first time dummy is dropped without of generality to avoid perfect collinearity. (2.1) is the individual-level population model.

A distinctive feature of the model is the incorporation of $G$ predetermined time-invariant groups. Denote the group membership of individual $i$ by $g_i$, a scalar random variable that takes on values in $\{1, \dots, G\}$. These $G$ groups form the entities to be tracked in the subsequently constructed pseudo panel. Additionally, these $G$ groups give rise to a more compact representation of (2.1). To see that, project $f_i$ onto the cohort level and define the cohort-specific fixed effects as

$$\alpha_g = E(f_i|g), g = 1, ..., G \tag{2.2}$$

where $E(\cdot|g)$ is short for $E(\cdot|g_i = g)$. If $g_i$ is not evaluated at $g$ but left as $g_i$ in the conditioning set, $\alpha_g$ becomes the random cohort effect $\alpha_{g_i}$, allowing us to define the individual-specific fixed effect net of the cohort effect as

$$e_i = f_i - \alpha_{g_i}. \tag{2.3}$$

Note that the exogeneity of $g_i$ with respect to $e_i$,

$$E(e_i|g) = 0, \tag{2.4}$$

holds by construction.[6] Let $\mathbf{c}_i \equiv (c_{i2}, ..., c_{iG})$ with $c_{ig} = 1_{\{g_i=g\}}$ be the $1 \times (G-1)$ cohort dummy vector and

[4] $1_{\{\cdot\}}$ is the indicator function that equals 1 if the condition in $\{\cdot\}$ is true and 0 otherwise.

[5] A parameter notation, such as $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$, denotes the true parameter value if not explicitly stated otherwise.

[6] (2.4) is not a an additional assumption; once (2.1) is given, it follows as a result of including a full set of non-redundant cohort dummies in the equivalent representation (2.5). See Imbens & Wooldridge (2007) for more.

$\boldsymbol{\alpha} \equiv (\alpha_2, \ldots, \alpha_G)'$ be the $(G-1) \times 1$ vector of cohort effects.[7] Then (2.1) can be rewritten compactly as

$$y_{it} = \underline{\mathbf{x}}_{it}\boldsymbol{\theta} + e_i + u_{it}, \; t = 1, \ldots, T \tag{2.5}$$

where, for $\underline{K} = K+G+T-2$, $\underline{\mathbf{x}}_{it} \equiv (\mathbf{x}_{it}, \mathbf{d}_t, \mathbf{c}_i)$ is the $1 \times \underline{K}$ extended vector of regressors and $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}', \boldsymbol{\eta}', \boldsymbol{\alpha}')'$ is the $\underline{K} \times 1$ vector of all structure parameters of interest. . For ease of use later, it is useful to define the individual-level composite error as

$$\varepsilon_{it} = e_i + u_{it} = (f_i - \alpha_{g_i}) + u_{it} = y_{it} - \underline{\mathbf{x}}_{it}\boldsymbol{\theta}. \tag{2.6}$$

$\varepsilon_{it}$ will be the vehicle in deriving the asymptotic theory later.

A key identification assumption imposed in pseudo panel models is the exogeneity of $g_i$ w.r.t. $u_{it}$, i.e.,

$$E(u_{it}|g) = 0, \; g = 1, 2, \ldots, G. \tag{2.7}$$

The significance of (2.7) is that it gives rise to the cohort-level population model. Specifically, taking the conditional expectation of (2.5) given $g$ leads to

$$\mu^y_{gt} = \boldsymbol{\mu}^{\underline{\mathbf{x}}}_{gt}\boldsymbol{\theta}, \; g = 1, \ldots, G, \; t = 1, \ldots, T, \tag{2.8}$$

where $\mu^y_{gt} \equiv E(y_{it}|g)$ and $\boldsymbol{\mu}^{\underline{\mathbf{x}}}_{gt} \equiv E(\underline{\mathbf{x}}_{it}|g)$ represent the cohort means of $y_{it}$ and $\underline{\mathbf{x}}_{it}$ in group $g$, respectively. Since the $G$ groups and $T$ time periods effectively divide the population over time into $GT$ group-time cells, $\mu^y_{gt}$ and $\boldsymbol{\mu}^{\underline{\mathbf{x}}}_{gt}$ can also be termed as the population cell means of $y_{it}$ and $\underline{\mathbf{x}}_{it}$ in group-time cell $(g, t)$. In the GMM framework in Inoue (2008), the conditions in (2.8) are known as the moment conditions. In the MD framework, they are often called the restrictions or constraints on the structural and reduced-form parameters. $\mu^y_{gt}$'s and $\boldsymbol{\mu}^{\underline{\mathbf{x}}}_{gt}$'s are the reduced-form parameters since they are intermediate parameters not of direct interest. Clearly, (2.8) has a panel structure if we treat each cohort as a research unit to follow over time, and hence leading to the term pseudo panel models.

It is worth noting that (2.4) and (2.7) together essentially indicates that $g_i$ is a valid instrumental variable (IV) w.r.t. the composite error $\varepsilon_{it}$. In addition, to identify $\boldsymbol{\theta}$ from (2.8), it is necessary for $g_i$ to be correlated with the non-constant variables in $\mathbf{x}_{it}$, and each entry of $\boldsymbol{\mu}^{\underline{\mathbf{x}}}_{gt}$ must exhibit sufficient variation across the $GT$ group-time cells. Therefore, $g_i$ is also a relevant IV, albeit with a stronger relevance assumption. It is noteworthy that $g_i$ is not used as an IV in the conventional way in pseudo panel estimation. Instead, it is

[7] The dimensions of $\mathbf{c}_i$ and $\boldsymbol{\alpha}$ are $G-1$ for the same perfect collinearity consideration as the dimensions of $\mathbf{d}_t$ and $\boldsymbol{\eta}$. Note that $\mathbf{c}_i$ is an equivalent reparameterization of the group membership variable $g_i$. Observing $\mathbf{c}_i$ would imply $g_i$ and vice versa.

used to project the individual-level model onto the cohort level to eliminate the composite error $\varepsilon_{it}$. But since $g_i$ is discrete, this projection preserves the same set of information as the conventional way of using IV.

As a final remark of this subsection, notice that the cohort-level composite residual in group-time cell $(g,t)$, defined as

$$\mu^{\varepsilon}_{gt} \equiv E(\varepsilon_{it}|g) = \mu^{y}_{gt} - \boldsymbol{\mu}^{\mathbf{x}}_{gt}\boldsymbol{\theta}, \tag{2.9}$$

satisfies $\mu^{\varepsilon}_{gt} = 0$ by the exogeneity of $g_i$ w.r.t. $\varepsilon_{it}$, or as a rephrasing of (2.8) for $g = 1, ..., G$ and $t = 1, ..., T$. The feasible version of (2.9) will serve as the tool to derive the optimal weighting matrix later on.

## 2.2 Minimum distance estimation of pseudo panel models

(2.8) suggests a straightforward way to estimate $\boldsymbol{\theta}$: Since $\mu^{y}_{gt}$ and $\boldsymbol{\mu}^{\mathbf{x}}_{gt}$ can be readily estimated by $\hat{\mu}^{y}_{gt}$ and $\hat{\boldsymbol{\mu}}^{\mathbf{x}}_{gt}$, the sample cohort means of $y_{it}$ and $\underline{\mathbf{x}}_{it}$, respectively, $\boldsymbol{\theta}$ can be estimated through a regression of $\hat{\mu}^{y}_{gt}$ on $\hat{\boldsymbol{\mu}}^{\mathbf{x}}_{gt}$, $g = 1, \ldots, G$, $t = 1, \ldots, T$, which leads to the following FE estimator applied to the constructed pseudo panel:

$$\check{\boldsymbol{\theta}} = \left( \sum_{g,t} \hat{\boldsymbol{\mu}}^{\mathbf{x}\prime}_{gt} \hat{\boldsymbol{\mu}}^{\mathbf{x}}_{gt} \right)^{-1} \sum_{g,t} \hat{\boldsymbol{\mu}}^{\mathbf{x}\prime}_{gt} \hat{\mu}^{y}_{gt}. \tag{2.10}$$

It is well known, however, that while $\check{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}$, the corresponding FE s.e. ignores the estimation errors in $\hat{\mu}^{y}_{gt}$ and $\hat{\boldsymbol{\mu}}^{\mathbf{x}}_{gt}$ and thus would result in invalid statistical inference. This holds true even when the inference is made robust to heteroskedasticity and/or serial correlation. In contrast, the MD approach addresses of the estimation errors in $\hat{\mu}^{y}_{gt}$ and $\hat{\boldsymbol{\mu}}^{\mathbf{x}}_{gt}$. Moreover, it can incorporate optimal weighting to utilize the $GT$ restrictions in (2.8) in the most asymptotically efficient manner. As demonstrated in Imbens & Wooldridge (2007) and subsequent sections, $\check{\boldsymbol{\theta}}$ is actually a particular MD estimator where the identity weighting matrix is used. Consequently, $\check{\boldsymbol{\theta}}$ is inefficient if the optimal weighting matrix deviates from the identity matrix.

To define the class of MD estimators, it is necessary to have an explicit account of the data set as it provides a tool to model the sampling design. The data set available in pseudo panel estimation typically comprises a series of cross-sectional samples from consecutive time periods. If we define $\mathbf{w}_{it} = (y_{it}, \mathbf{x}_{it})$ and let $\mathcal{I}_t$ be the index set for sample $t$, $t = 1, ..., T$, we can represent the $T$ cross sections in the data set as

$$\{(\mathbf{w}_{it}, g_i) : i \in \mathcal{I}_t, t = 1, \ldots, T\}. \tag{2.11}$$

Note that using a time-dependent index set $\mathcal{I}_t$ allows the same $i$ to consistently refer to the same individual in the population across different time periods. This treatment differs from that in Imbens & Wooldridge (2007), where $i$ is employed to index random draws, resulting in the same index $i$ typically representing different individuals at different time periods. While that treatment suffices for serially independent sampling, it lacks the sophistication required for the partially matched sampling considered here.

Explicitly modeling the data set also enables us to explicitly define $\hat{\mu}_{gt}^{\mathbf{w}}$, the sample cell mean of $\mathbf{w}_{it}$ in the group-time cell $(g,t)$. To see that, first define the random dummy indicator for the group membership of individual $i$ at time $t$ as

$$r_{itg} = 1_{\{g_i = g, i \in \mathcal{I}_t\}} \tag{2.12}$$

The randomness of $r_{itg}$ comes from the sampling procedure. Note that $r_{itg}$ depends on $t$ because $i \in \mathcal{I}_t$. Then $\hat{\mu}_{gt}^{\mathbf{w}}$ can be written as

$$\hat{\mu}_{gt}^{\mathbf{w}} = n_{gt}^{-1} \sum_{i \in \mathcal{I}_t} r_{itg} \mathbf{w}_{it}. \tag{2.13}$$

where $n_{gt} = \sum_{i \in \mathcal{I}_t} r_{itg}$ is the sample size of cell $(g,t)$, properly treated as a random variable. Other sample cohort means, such as $\hat{\mu}_{gt}^{y}$,$\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{x}}$ and $\hat{\mu}_{gt}^{\varepsilon}$, can be defined similarly. These sample cohort means are consistent for their corresponding population cohort means under weak regularity conditions.

The MD approach recovers the asymptotics of the structural parameters from those of the reduced-form parameters. Therefore, it is also useful to have a representation of the joint asymptotic distribution of $\hat{\mu}_{gt}^{\mathbf{w}}$ for all $g$ and $t$. For this purpose, define

$$\boldsymbol{\pi} \equiv (\boldsymbol{\mu}_{11}^{\mathbf{w}}, \boldsymbol{\mu}_{12}^{\mathbf{w}}, \ldots, \boldsymbol{\mu}_{GT}^{\mathbf{w}})' \tag{2.14}$$

as the $(K+1)GT \times 1$ reduced-form parameter vector that collects all the population cohort means in one column. The natural estimator of $\boldsymbol{\pi}$, denoted by $\hat{\boldsymbol{\pi}}$, can be obtained by substituting $\boldsymbol{\mu}_{gt}^{\mathbf{w}}$ with $\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{w}}$ in (2.14). Let $n = \sum_{t=1}^{T} n_t$ be the total sample size of all cross sections. Then under fairly weak regularity conditions,

$$\sqrt{n}\,(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Omega}) \tag{2.15}$$

where $\boldsymbol{\Omega}$ is the $GT(K+1) \times GT(K+1)$ asymptotic variance. The structure of $\boldsymbol{\Omega}$ depends on assumptions on the population (finite v.s. infinite population) and the sampling design (serially independent sampling v.s. partially matched sampling), and the latter is the focus of this paper. Without going into details, $\boldsymbol{\Omega}$ under serially independent sampling can be written as

$$\boldsymbol{\Omega} = \mathbf{diag}\left\{(\rho_g \kappa_t)^{-1} \boldsymbol{\Omega}_{gt}^{\mathbf{w}}\right\} \tag{2.16}$$

where $\boldsymbol{\Omega}_{gt}^{\mathbf{w}} \equiv Var(\mathbf{w}_{it}|g)$ is the variance of $\mathbf{w}_{it}$ in cell $(g,t)$, $\rho_g = P(r_{itg} = 1)$ is the fraction of the population in group $g$ at time $t$[8], and $\kappa_t$ is the fraction of all observations accounted for by cross section $t$. $\rho_g$ can be consistently estimated by$\hat{\rho}_{gt} \equiv n_{gt}/n_t \xrightarrow{p} \rho_g$,whereas $\kappa_t$ can be consistently estimated by $\hat{\kappa}_t \equiv n_t/n \xrightarrow{p} \kappa_t$.[9]

The set of equations in (2.8) are the restrictions that the MD approach explores in pseudo panel models to identify the structural parameters. In a general MD framework, we can write the restrictions that links the structural parameter vector $\boldsymbol{\theta}$ to the reduced-form parameter vector $\boldsymbol{\pi}$ for a generic pair $(\boldsymbol{\pi}, \boldsymbol{\theta})$ as $\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta}) = 0$ where $\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta})$ is a $J \times 1$ vector-valued function. For the restrictions in (2.8), $J = GT$ and the particular $\mathbf{h}$ function is

$$\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta}) = -\boldsymbol{\mu}^{y} + \boldsymbol{\mu}^{\mathbf{x}}\boldsymbol{\theta} \tag{2.17}$$

where $\boldsymbol{\mu}^{y} \equiv (\mu_{11}^{y}, \mu_{12}^{y}, \ldots, \mu_{GT}^{y})'$ is a $GT \times 1$ vector and $\boldsymbol{\mu}^{\mathbf{x}} \equiv (\boldsymbol{\mu}_{11}^{\mathbf{x}}{}', \boldsymbol{\mu}_{12}^{\mathbf{x}}{}', \ldots \boldsymbol{\mu}_{GT}^{\mathbf{x}}{}')$ is a $GT \times \underline{K}$ matrix. (2.17) can be viewed as a residual function because it generates the negative of the cohort-level composite residuals $\mu_{gt}^{\varepsilon}$'s if $\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta})$ is evaluated at the true parameter value. In fact, if we define $\boldsymbol{\mu}^{\varepsilon} = (\mu_{11}^{\varepsilon}, \mu_{12}^{\varepsilon}, \ldots, \mu_{GT}^{\varepsilon})'$, it follows from $\mu_{gt}^{\varepsilon} = 0$ that, when evaluated at the true parameter value,

$$\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta}) = -\boldsymbol{\mu}^{\varepsilon} = 0, \tag{2.18}$$

Note that often $GT > \underline{K}$ in pseudo panel models, so the system is usually over-identified for $\boldsymbol{\theta}$, which merits the use of some positive semi-definite weighting matrix. Depending on the choice of the weighting matrix, we can have the FE estimator, the optimal MD pseudo panel estimator and so on.

To construct the optimal MD estimator, a representation of the optimal weighting matrix is also needed. The optimal weighting matrix in pseudo panel MD estimation is closely related to $\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta})$. In fact, given (2.18), it can be shown that, under standard regularity conditions (see Newey & McFadden 1994), the asymptotic distribution of $\sqrt{n}\left[\mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}) - \mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta})\right]$ can be written as

$$\sqrt{n}\mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{M}(\boldsymbol{\theta})\right). \tag{2.19}$$

where $\mathbf{M}(\boldsymbol{\theta})$ is the $GT \times GT$ asymptotic variance of $\sqrt{n}\mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta})$ and $\boldsymbol{\theta}$ is the true parameter value. Intuitively, $\mathbf{M}(\boldsymbol{\theta})^{-1}$ should be the optimal weighting matrix in MD estimation because it standardizes the asymptotic variance to an identity matrix if $\sqrt{n}\mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta})$ is multiplied by the square root of $\mathbf{M}(\boldsymbol{\theta})^{-1}$ (see Kodde et al. 1990, Newey & McFadden 1994 for formal proofs). Some algebra shows that

$$\mathbf{M}(\boldsymbol{\theta}) = \mathbf{diag}\left\{(\rho_g \kappa_t)^{-1} Var\left(\varepsilon_{it}|g\right)\right\} \tag{2.20}$$

[8] In general, we can use $\rho_{gt} = P(r_{itg} = 1)$ which varies over cohort and time. To simplify the discussion, however, $\rho_{gt}$ is assumed to be constant across $t$, i.e., $\rho_{gt} = \rho_g$ hereafter, which is the case if a stable population is assumed.

[9] Details on deriving the $\boldsymbol{\Omega}$ in (2.16) and its consistent estimator $\hat{\boldsymbol{\Omega}}$ will be given in the online appendix. The structure of $\boldsymbol{\Omega}$ under partially matched sampling will be provided in Section 3.

which is the $GT \times GT$ diagonal matrix with the within-cell variances of $\varepsilon_{it}$ weighted by the inverse of the relative cell sizes $\rho_{gt}\kappa_t$ on the diagonal. Hence, a more convenient estimator for $\mathbf{M}(\boldsymbol{\theta})$ is

$$\hat{\mathbf{M}} = \mathbf{diag}\left\{(n_{gt}/n)^{-1}\hat{\sigma}^2_{\varepsilon,gt}\right\}$$

where $\hat{\sigma}^2_{\varepsilon,gt}$ is the sample variance of $\varepsilon_{it}$ within cell $(g, t)$ and $n_{gt}/n$ is a consistent estimator for $\rho_g\kappa_t$. An explicit expression of $\hat{\sigma}^2_{\varepsilon,gt}$ is given in (3.13) later.

With all the groundwork laid out, the feasible optimal MD pseudo panel estimator for $\boldsymbol{\theta}$ can now be formulated as

$$\hat{\boldsymbol{\theta}}^{opt} = \underset{\boldsymbol{\theta}}{argmin}\ \mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta})'\hat{\mathbf{M}}^{-1}\mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}). \tag{2.21}$$

That is, $\hat{\boldsymbol{\theta}}^{opt}$ is the minimizer of $\mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta})'\hat{\mathbf{M}}^{-1}\mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta})$, a quadratic Euclidean distance function. The first-order condition to the optimization problem (2.21) is $\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}\,\prime}\hat{\mathbf{M}}^{-1}(-\boldsymbol{\mu}^{y} + \boldsymbol{\mu}^{\underline{\mathbf{x}}}\boldsymbol{\theta}) = \mathbf{0}$; solving yields

$$\hat{\boldsymbol{\theta}}^{opt} = \left[\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}\,\prime}\hat{\mathbf{M}}^{-1}\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}}\right]^{-1}\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}\,\prime}\hat{\mathbf{M}}^{-1}\hat{\boldsymbol{\mu}}^{y}. \tag{2.22}$$

(2.22) is a GLS estimator performed on the cohort means with the weighting matrix $\hat{\mathbf{M}}^{-1}$. It is also clear from (2.22) that if the identity weighting matrix is used in replacement of $\hat{\mathbf{M}}^{-1}$, the FE estimator $\check{\boldsymbol{\theta}}$ defined in (2.10) follows.

Without going into details, it can be shown that the asymptotic variance of $\hat{\boldsymbol{\theta}}^{opt}$ is given by (see Kodde et al. 1990, Imbens & Wooldridge 2007, Inoue 2008 and Jia 2019)

$$Avar\left(\hat{\boldsymbol{\theta}}^{opt}\right) = n^{-1}\left[\boldsymbol{\mu}^{\underline{\mathbf{x}}\,\prime}\mathbf{M}^{-1}\boldsymbol{\mu}^{\underline{\mathbf{x}}}\right]^{-1} \tag{2.23}$$

for which an estimator follows by replacing $\boldsymbol{\mu}^{\underline{\mathbf{x}}}$ and $\mathbf{M}$ with their respective estimators $\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}}$ and $\hat{\mathbf{M}}$, i.e.,

$$\widehat{Avar\left(\hat{\boldsymbol{\theta}}^{opt}\right)} = n^{-1}\left[\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}\,\prime}\hat{\mathbf{M}}^{-1}\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}}\right]^{-1} \tag{2.24}$$

$\check{\boldsymbol{\theta}}$ is not the focus of the present paper so its asymptotic variance is skipped .

# 3 The partially matched sample correction

Compared to the standard pseudo panel model under the serially independent sampling in the last section, a major extension in this paper is to allow for nonzero asymptotic covariance between $\sqrt{n}(\hat{\mu}^{\mathbf{w}}_{gt} - \mu^{\mathbf{w}}_{gt})$ and $\sqrt{n}(\hat{\mu}^{\mathbf{w}}_{gs} - \mu^{\mathbf{w}}_{gs})$ for $t \neq s$, which in turn means that it allows for nonzero asymptotic covariance between

$\sqrt{n}\hat{\mu}^{\varepsilon}_{gt}$ and $\sqrt{n}\hat{\mu}^{\varepsilon}_{gs}$ (i.e., the $[(g-1)t+T]$-th and $[(g-1)s+T]$-th rows of $\sqrt{n}\mathbf{h}(\hat{\boldsymbol{\pi}},\boldsymbol{\theta})$). Note that, because of random sampling within each cross section, $\sqrt{n}(\hat{\mu}^{\mathbf{w}}_{gt}-\mu^{\mathbf{w}}_{gt})$ and $\sqrt{n}(\hat{\mu}^{\mathbf{w}}_{lt}-\mu^{\mathbf{w}}_{lt})$ for $g \neq l$ are still asymptotically uncorrelated.

### 3.1 The partially matched sampling design

To derive the joint distribution of $\sqrt{n}(\hat{\mu}^{\mathbf{w}}_{gt}-\mu^{\mathbf{w}}_{gt})$ and $\sqrt{n}(\hat{\mu}^{\mathbf{w}}_{gs}-\mu^{\mathbf{w}}_{gs})$, it helps to sort through the sampling procedure first. This section uses the MORG files extracted from CPS as the motivating example. The simplified sampling procedure distilled of the MORG example, however, applies more generally to any data set that fits the simplified sampling procedure.

In CPS, new households enter each month, and every household is interviewed for 4 consecutive months, ignored for the next 8 months, and then interviewed again for another 4 month before permanently exists the survey (CPS Technical Documentation 2014). The MORG files are a series of yearly data sets that includes only the data obtained at the 4th and 8th interviews for each household because extra questions regarding hours/earning are asked only at those extracts. Although the special rotation group design is the underlying sampling procedure truly used in generating the MORG files, modeling details up to the rotation group design is unnecessary if we plan to use the MORG files as annual cross-sectional samples. Because of the particular rotation group design, approximately half of the observations in the MORG sample at a given year are matched with half of those in the subsequent year. Therefore, the MORG sampling design can be equivalently simplified as follows. For simplicity, assume a super population with infinite size exists.

**Definition 1.** Simplified Partially Matched Sampling Design

Suppose there are $T$ time periods and $t = 1, ..., T$, $n_t$ is the sample size of cross section $t$, and $n_{t,stay}$ is the number of observations to keep to the next period. Then the $T$ partially matched cross sectional sampled are obtained through the following steps:

1. At time $t = 1$: randomly sample $n_1$ observations from the population to form sample 1. Randomly select a subsample of $n_{1,stay}$ observations from sample 1 to keep to period 2.
2. At time $t$ for $t > 1$: Randomly sample $(n_t - n_{t-1,stay})$ observations from the population. Combine with the subsample left from $t-1$ to get sample $t$. Stop here if $t = T$. If $t < T$, randomly select a subsample of $n_{t,stay}$ observations in the newly drawn subsample of $(n_s - n_{t,stay})$ observations to keep to period $t+1$. Repeat step 2 until stop.

A few remarks regarding the sampling procedure follow. First, although the sampling procedure is motivated by MORG, it applies to other sampling designs that lead to the same simplified procedure. The monthly Australia Labour Force Survey (LFS) is such an example.

Second, although $n_{t,stay}$ in general depends on $t$, in relevant examples it usually is a fixed number. In MORG, $n_{t,stay} = n_t/2$, which implies $m_{t,t+1} = n_t/2$ and $m_{ts} = 0$ for any $s > t+1$. In the simulation study later, we use a constant number for $n_t$ and a fixed matching rate between adjacent samples. As a result, $n_{t,stay}$ and $m_{t,t+1}$ are constants as well in the simulation study.

Third, it is worth noting how to randomly select a given fixed number of observations out of a given sample. Suppose the sample size is 4 and we want to randomly sample half of them. In this scenario, there are $C_4^2 = 4!/[2!(4-2)!] = 6$ different ways to achieve the gold. To ensure randomness, we can throw an even six-sided dice to randomly select exactly half of the sample. This idea can be easily generalized to situations where we want to randomly select a subsample of a fixed size out of a sample of size $n$. A relevant question is whether the matched subsample generated in this way is independent of the unmatched subsample. The answer is affirmative, provided that the strategy described here is used. It's important to distinguish this approach from another where we repeatedly toss a fair coin for each individual in the sample to decide if they should be included in the selected subsample. In the latter scenario, if the sample size is 4, there are $2^2$ possibilities, and the resulting subsample size could be any number in $\{0, 1, 2, 3, 4\}$. While both strategies aim to select half of the given sample conceptually, the realized subsample size itself is random in the latter strategy. Either strategy could be useful depending on the application. In the MORG example, the former seems more plausible, considering that "in any given month, one-eighth of the housing units are interviewed for the first month" according to CPS. Admittedly, this is not determinant evidence that the former strategy is adopted in CPS. However, it seems reasonable to assume that, in a steady state, the same number of new households are drawn each month in CPS. Moreover, the former strategy makes the asymptotic analysis a bit easier.

## 3.2 Asymptotics in the presence of matched subsamples among samples from different periods

With the sampling design well defined, we can now derive the asymptotics under partially matched sampling. For $t < s$, Let $\mathcal{I}_{ts}$ be the index set for the matched subsample between sample $t$ and sample $s$. Note that $\mathcal{I}_{ts} = \mathcal{I}_t \cap \mathcal{I}_s$. Denote the size of $\mathcal{I}_{ts}$ by $m_{ts}$, i.e., $m_{ts} = |\mathcal{I}_{ts}|$. In MORG, for instance, only adjacent samples share matched subsamples with $n_{t,stay} = n_t/2$, which implies $m_{t,t+1} = n_t/2$ and $m_{ts} = 0$ for any $s > t+1$. With $\mathcal{I}_{ts}$, the sample cohort mean $\hat{\mu}_{gt}^{\mathbf{w}}$ for a given pair $(t, s)$ where $t < s$ can be decomposed into the sum of the sample cohort mean of the unmatched subsample and that of the the matched, i.e.,

$$\hat{\mu}_{gt}^{\mathbf{w}} = n_{gt}^{-1} \sum_{i \in \mathcal{I}_t} r_{itg} \mathbf{w}_{it} = n_{gt}^{-1} \sum_{i \in \mathcal{I}_{ts}} r_{itg} \mathbf{w}_{it} + n_{gt}^{-1} \sum_{i \in \mathcal{I}_t \setminus \mathcal{I}_{ts}} r_{itg} \mathbf{w}_{it} \tag{3.1}$$

where $\mathcal{I}_t \backslash \mathcal{I}_{ts}$ is the difference set. A similar decomposition exists for $\hat{\mu}^{\mathbf{w}}_{gs}$ for the same pair $(t, s)$, which can be obtained by replacing $t$ with $s$ and $s$ with $t$ in (3.1) and noticing that $\mathcal{I}_{ts} = \mathcal{I}_{st}$. Because of the sampling design, the matched and unmatched subsamples in a given period are independent; the unmatched subsample in period $t$ is also independent of that in period $s$. Hence, we can separately derive the asymptotic distribution of $n_{gt}^{-1} \sum_{i \in \mathcal{I}_t \backslash \mathcal{I}_{ts}} r_{itg} \mathbf{w}_{it}$, the asymptotic distribution of $n_{gs}^{-1} \sum_{i \in \mathcal{I}_s \backslash \mathcal{I}_{ts}} r_{isg} \mathbf{w}_{it}$, and the asymptotic joint distribution of $n_{gt}^{-1} \sum_{i \in \mathcal{I}_{ts}} r_{itg} \mathbf{w}_{it}$ and $n_{gs}^{-1} \sum_{i \in \mathcal{I}_{ts}} r_{isg} \mathbf{w}_{is}$ and then put them together to get the joint asymptotic distribution of $\sqrt{n}(\hat{\mu}^{\mathbf{w}}_{gt} - \mu^{\mathbf{w}}_{gt})$ and $\sqrt{n}(\hat{\mu}^{\mathbf{w}}_{gs} - \mu^{\mathbf{w}}_{gs})$. This result is summarized in the following lemma.

**Lemma 1.** *Suppose the population DGP is given by (2.1) and the sampling design is defined as in Definition 1. In addition, assume $\rho_{gt} = \rho_g$ is a constant for ease of illustration. Let $\psi_{ts,t}$ be the fraction of sample $t$ that is in the matched subsample between sample $t$ and sample $s$ and $\psi_{ts,s}$ be defined similarly. Let $\mathbf{\Omega}^{\mathbf{w}}_{gts} = Cov(\mathbf{w}_{it}, \mathbf{w}_{is} | g, i \in \mathcal{I}_{ts})$, and define $\Psi^{\mathbf{w}}_{gts} = (\psi_{ts,t}\psi_{ts,s})^{\frac{1}{2}\mathbb{I}_{ts}} (\rho_g \kappa_t \rho_g \kappa_s)^{-\frac{1}{2}} \mathbf{\Omega}^{\mathbf{w}}_{gts}$. Then,*

$$\sqrt{n}\left[\begin{pmatrix} \hat{\mu}^{\mathbf{w}}_{gt} \\ \hat{\mu}^{\mathbf{w}}_{gs} \end{pmatrix} - \begin{pmatrix} \mu^{\mathbf{w}}_{gt} \\ \mu^{\mathbf{w}}_{gs} \end{pmatrix}\right] \xrightarrow{d} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Psi^{\mathbf{w}}_{gtt} & \Psi^{\mathbf{w}}_{gst} \\ \Psi^{\mathbf{w}}_{gts} & \Psi^{\mathbf{w}}_{gss} \end{pmatrix}\right] \tag{3.2}$$

*Proof.* Let $z_{itg} = r_{itg}(\mathbf{w}_{it} - \mu^{\mathbf{w}}_{gt})$, $\bar{z}_{0tsg} = (n_t - m_{ts})^{-1} \sum_{i \in \mathcal{I}_t \backslash \mathcal{I}_{ts}} z_{itg}$ and $\bar{z}_{1tsg} = m_{ts}^{-1} \sum_{i \in \mathcal{I}_{ts}} z_{itg}$. It can be shown that $E(z_{itg}) = 0$ and $Var(z_{itg}) = \rho_g \mathbf{\Omega}^{\mathbf{w}}_{gt}$ where $\mathbf{\Omega}^{\mathbf{w}}_{gt} = Var(\mathbf{w}_{it}|g)$ is as defined in previous sections. Let $\hat{\kappa}_t = n_t/n$, $\hat{\rho}_{gt} = n_{gt}/n_t$, $\hat{\psi}_{ts,t} = m_{ts}/n_t$ and $\hat{\psi}_{ts,s} = m_{ts}/n_s$. Using (3.1) to write

$$\sqrt{n}(\hat{\mu}^{\mathbf{w}}_{gt} - \mu^{\mathbf{w}}_{gt}) = \hat{\kappa}_t^{-1/2} \hat{\rho}_{gt}^{-1} \hat{\psi}_{ts,t}^{1/2} \cdot \sqrt{m_{ts}} \bar{z}_{1tsg} + \hat{\kappa}_t^{-1/2} \hat{\rho}_{gt}^{-1} (1 - \hat{\psi}_{ts,t})^{1/2} \cdot \sqrt{n_t - m_{ts}} \bar{z}_{0tsg} \tag{3.3}$$

Notice that $\sqrt{m_{ts}} \bar{z}_{1tsg} \xrightarrow{d} N\left(0, \rho_g \mathbf{\Omega}^{\mathbf{w}}_{gt}\right)$ and $\sqrt{n_t - m_{ts}} \bar{z}_{0tsg} \xrightarrow{d} N\left(0, \rho_g \mathbf{\Omega}^{\mathbf{w}}_{gt}\right)$ by the usual Central Limit Theorem (CLT). Notice also that $\hat{\kappa}_t \to \kappa_t$, $\hat{\rho}_{gt} \to \rho_g$ and $\hat{\psi}_{ts,t} \to \psi_{ts,t}$. Then, by the asymptotic equivalence lemma in Wooldridge (2010) and the fact that the $z_{itg}$'s in $\mathcal{I}_t \backslash \mathcal{I}_{ts}$ are independent of those in $\mathcal{I}_{ts}$, the asymptotic distributions of the two terms in (3.3) can simply be added up to get

$$\sqrt{n}(\hat{\mu}^{\mathbf{w}}_{gt} - \mu^{\mathbf{w}}_{gt}) \xrightarrow{d} N\left[0, (\kappa_t \rho_g)^{-1} \mathbf{\Omega}^{\mathbf{w}}_{gt}\right]$$

A similar result holds for $\hat{\mu}^{\mathbf{w}}_{gs}$.

To derive the asymptotic joint distribution of $\sqrt{n}(\hat{\mu}^{\mathbf{w}}_{gt} - \mu^{\mathbf{w}}_{gt})$ and $\sqrt{n}(\hat{\mu}^{\mathbf{w}}_{gs} - \mu^{\mathbf{w}}_{gs})$, notice that $m_{ts} = m_{st}$ so that $\bar{z}_{0stg} = (n_s - m_{ts})^{-1} \sum_{i \in \mathcal{I}_s \backslash \mathcal{I}_{ts}} z_{isg}$ and $\bar{z}_{1stg} = m_{ts}^{-1} \sum_{i \in \mathcal{I}_{ts}} z_{isg}$. Note that $\bar{z}_{1stg}$ and $\bar{z}_{1tsg}$ different because $r_{itg}$ and $r_{its}$ are. Write

$$\sqrt{n}(\hat{\mu}^{\mathbf{w}}_{gs} - \mu^{\mathbf{w}}_{gs}) = \hat{\kappa}_s^{-1/2}\hat{\rho}_{gs}^{-1}\hat{\psi}_{ts,s}^{1/2} \cdot \sqrt{m_{ts}}\bar{z}_{1stg} + \hat{\kappa}_s^{-1/2}\hat{\rho}_{gs}^{-1}(1-\hat{\psi}_{ts,s})^{1/2} \cdot \sqrt{n_t - m_{ts}}\bar{z}_{0stg}$$

Since the matched subsample is a subset randomly selected from sample $t$ by the sampling design, it can be shown that $E\left(z_{itg}|i \in \mathcal{I}_{ts}\right) = E\left(z_{isg}|i \in \mathcal{I}_{ts}\right) = 0$ and $Cov\left(z_{itg}, z_{isg}|i \in \mathcal{I}_{ts}\right) = \rho_g \mathbf{\Omega}^{\mathbf{w}}_{gts}$. Then, by the usual CLT,

$$\begin{bmatrix} \sqrt{m_{ts}}\bar{z}_{1tsg} \\ \sqrt{m_{ts}}\bar{z}_{1stg} \end{bmatrix} \xrightarrow{d} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \rho_g \begin{pmatrix} \mathbf{\Omega}^{\mathbf{w}}_{gt} & \mathbf{\Omega}^{\mathbf{w}}_{gts} \\ \mathbf{\Omega}^{\mathbf{w}}_{gts} & \mathbf{\Omega}^{\mathbf{w}}_{gs} \end{pmatrix}\right]$$

By the sampling design, the $z_{isg}$'s in $\mathcal{I}_s \backslash \mathcal{I}_{ts}$ is independent of the $z_{itg}$'s in $\mathcal{I}_t \backslash \mathcal{I}_{ts}$. Hence, by CLT,

$$\begin{bmatrix} \sqrt{n_t - m_{ts}}\bar{z}_{0tsg} \\ \sqrt{n_s - m_{ts}}\bar{z}_{0stg} \end{bmatrix} \xrightarrow{d} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \rho_g \begin{pmatrix} \mathbf{\Omega}^{\mathbf{w}}_{gt} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega}^{\mathbf{w}}_{gs} \end{pmatrix}\right]$$

Combining all above and applying the asymptotic equivalence lemma again lead to (3.2). □

It is worth emphasizing why the decomposition in (3.1) is necessary to derive the asymptotics in the presence of a matched subsample between sample $t$ and sample $s$. With serially independent sampling, $\hat{\mu}^{\mathbf{w}}_{gt}$ and $\hat{\mu}^{\mathbf{w}}_{gs}$ are independent of each other, and we can derive their asymptotics separately and then join them together as in Imbens & Wooldridge (2007). With partially matched sampling, this trick no longer works as $\hat{\mu}^{\mathbf{w}}_{gt}$ and $\hat{\mu}^{\mathbf{w}}_{gs}$ are dependent. The piece-wise asymptotic analysis in the proof of Lemma 1 based on the decomposition in (3.1) solves the complication cause by the dependence between $\hat{\mu}^{\mathbf{w}}_{gt}$ and $\hat{\mu}^{\mathbf{w}}_{gs}$.

Lemma 1 implies that we can write the asymptotic distribution of $\sqrt{n}(\hat{\mu}^{\mathbf{w}}_{gt} - \mu^{\mathbf{w}}_{gt})$ for $t = 1, ..., T$ jointly as

$$\sqrt{n}(\hat{\boldsymbol{\mu}}^{\mathbf{w}}_{g} - \boldsymbol{\mu}^{\mathbf{w}}_{g}) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{\Psi}^{\mathbf{w}}_{g}\right), g = 1, ..., G, \tag{3.4}$$

where $\hat{\boldsymbol{\mu}}^{\mathbf{w}}_{g} = (\hat{\boldsymbol{\mu}}^{\mathbf{w}}_{g1}, ..., \hat{\boldsymbol{\mu}}^{\mathbf{w}}_{gT})$, $\boldsymbol{\mu}^{\mathbf{w}}_{g}$ is similarly defined, and $\mathbf{\Psi}^{\mathbf{w}}_{g}$ is the square matrix with $(\kappa_t\rho_g)^{-1}\mathbf{\Omega}^{\mathbf{w}}_{gt}$ on the $t$-th principal diagonal block and $(\psi_{ts,t}\psi_{ts,s})^{1/2}(\kappa_t\kappa_s)^{-1/2}\rho_g^{-1}\mathbf{\Omega}^{\mathbf{w}}_{gts}$ on the $(t,s)$-th block for $t \neq s$. That is,

$$\mathbf{\Psi}^{\mathbf{w}}_{g} = \left\{(\psi_{ts,t}\psi_{ts,s})^{\frac{1}{2}\mathbb{I}_{ts}}(\rho_g\kappa_t\rho_g\kappa_s)^{-\frac{1}{2}}\mathbf{\Omega}^{\mathbf{w}}_{gts}\right\}_{TT}$$

where $\mathbb{I}_{ts} = 1_{\{t \neq s\}}$ is the indicator that equals 1 if $t \neq s$ and 0 otherwise. In turn, by noticing that $\boldsymbol{\pi} = (\boldsymbol{\mu}^{\mathbf{w}}_{1}, ..., \boldsymbol{\mu}^{\mathbf{w}}_{G})'$, stacking all the $G$ pieces together implies that we can write the joint asymptotic distribution of $\sqrt{n}(\hat{\mu}^{\mathbf{w}}_{gt} - \mu^{\mathbf{w}}_{gt})$ for $g = 1, ..., G$ and $t = 1, ..., T$ as

$$\sqrt{n}\left(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\right) \overset{d}{\to} N\left(\mathbf{0}, \boldsymbol{\Psi}^{\mathbf{w}}\right) \tag{3.5}$$

where $\boldsymbol{\Psi}^{\mathbf{w}} \equiv diag\left\{\boldsymbol{\Psi}^{\mathbf{w}}_g\right\}_G$ is the block diagonal matrix with $\boldsymbol{\Psi}^{\mathbf{w}}_g$ on the $g$-th block and 0 elsewhere. The asymptotics in (3.2), (3.4) and (3.5) are generalizations of the standard results under serially independent sampling to the case of partially matched sampling. When there is no matching, $\psi_{ts,t} = \psi_{ts,s} = 0$ and $\boldsymbol{\Psi}^{\mathbf{w}}$ degenerates to $\boldsymbol{\Omega}$ in (2.16).

The same asymptotic analysis can be performed on $\hat{\mu}^{\varepsilon}_{gt}$ where $\varepsilon_{it}$ is composite error term as defined in (2.6) . Notice that a similar decomposition to (3.1) exists for $\hat{\mu}^{\varepsilon}_{gt}$:

$$\hat{\mu}^{\varepsilon}_{gt} = n^{-1}_{gt} \sum_{i \in \mathcal{I}_{ts}} r_{it,g}\varepsilon_{it} + n^{-1}_{gt} \sum_{i \in \mathcal{I}_{ts}} r_{it,g}\varepsilon_{it}. \tag{3.6}$$

Notice also that $\mu^{\varepsilon}_{gt}$ is, for $j = (g-1)T + t$, the $j$-th row of $\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta})$ as in (2.17) with a negative sign added, and that $\sqrt{n}\hat{\mu}^{\varepsilon}_{gt}$ is the $j$-th row of $\sqrt{n}\mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta})$ as in (2.19) with a negative sign added. Hence, the asymptotic variance of the joint distribution of $\sqrt{n}\hat{\mu}^{\varepsilon}_{gt}$ and $\sqrt{n}\hat{\mu}^{\varepsilon}_{gs}$ essential determines (the inverse of) the optimal weighting matrix. The asymptotics regarding $\sqrt{n}\hat{\mu}^{\varepsilon}_{gt}$ is summarized in the following theorem.

**Theorem 1.** *Define* $\Psi^{\varepsilon}_{gts} = \left(\psi_{ts,t}\psi_{ts,s}\right)^{\frac{1}{2}\mathbb{I}_{ts}}\left(\rho_g\kappa_t\rho_g\kappa_s\right)^{-\frac{1}{2}}\sigma_{\varepsilon,gts}$ *where* $\mathbb{I}_{ts} \equiv 1_{\{t \neq s\}}$, $\sigma^2_{\varepsilon,g} \equiv Var(\varepsilon_{it}|g)$ *and* $\sigma_{\varepsilon,gts} \equiv Cov(\varepsilon_{it}, \varepsilon_{is}|g)$ *which degenerates to* $\sigma^2_{\varepsilon,g}$ *if* $t = s$. *Then, under the same DGP and sampling design in Lemma 1, the joint asymptotic distribution of* $\sqrt{n}\hat{\mu}^{\varepsilon}_{gt}$ *and* $\sqrt{n}\hat{\mu}^{\varepsilon}_{gs}$ *is*

$$\sqrt{n}\begin{pmatrix}\hat{\mu}^{\varepsilon}_{gt} \\ \hat{\mu}^{\varepsilon}_{gs}\end{pmatrix} \overset{d}{\to} N\left[\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\Psi^{\varepsilon}_{gtt} & \Psi^{\varepsilon}_{gst} \\ \Psi^{\varepsilon}_{gts} & \Psi^{\varepsilon}_{gss}\end{pmatrix}\right] \tag{3.7}$$

*In addition, the joint asymptotic distribution of* $\sqrt{n}\hat{\mu}^{\varepsilon}_{gt}$ *for* $g = 1, ..., G$ *and* $t = 1, ..., T$ *can be written as*

$$\sqrt{n}\hat{\boldsymbol{\mu}}^{\varepsilon} \overset{d}{\to} N\left(\mathbf{0}, \boldsymbol{\Psi}^{\varepsilon}\right) \tag{3.8}$$

*where* $\boldsymbol{\Psi}^{\varepsilon} = diag\left(\boldsymbol{\Psi}^{\varepsilon}_g\right)_G$ *is the block diagonal matrix with* $\boldsymbol{\Psi}^{\varepsilon}_g$ *on its g-th diagonal block for* $g = 1, ..., G$ *and* $\boldsymbol{\Psi}^{\varepsilon}_g$ *is the square matrix with* $(\kappa_t\rho_g)^{-1}\sigma^2_{\varepsilon,g}$ *on the t-th diagonal block and* $(\psi_{ts,t}\psi_{ts,s})^{1/2}(\kappa_t\kappa_s)^{-1/2}\rho_g^{-1}\sigma_{\varepsilon,gts}$ *on the* $(t,s)$*-th block for* $t \neq s$, *i.e.,*

$$\boldsymbol{\Psi}^{\varepsilon}_g = \left\{\Psi^{\varepsilon}_{gts}\right\}_{TT}. \tag{3.9}$$

*Proof.* Notice that for $\mathbf{w}_{it} = (y_{it}, \mathbf{x}_{it})$, $\varepsilon_{it} = y_{it} - \mathbf{x}_{it}\boldsymbol{\beta} - \alpha_{g_{it}} = (y_{it}, \mathbf{x}_{it})(1, -\boldsymbol{\beta}')' - \alpha_{g_{it}}$. Hence, $\hat{\mu}^{\varepsilon}_{gt} =$

$\hat{\mu}^{\mathbf{w}}_{gt}(1, -\boldsymbol{\beta}')' - \alpha_g$ and $\mu^{\varepsilon}_{gt} = \mu^{\mathbf{w}}_{gt}(1, -\boldsymbol{\beta}')' - \alpha_g$. Then the results in (3.7) and (3.8) follow from Lemma 1. $\square$

The joint asymptotic distribution of $\sqrt{n}\hat{\mu}^{\varepsilon}_{gt}$ and $\sqrt{n}\hat{\mu}^{\varepsilon}_{gs}$ in (3.7) shows why the off-diagonal elements in their asymptotic variance is nonzero. Let $\sigma^2_{f,g} = Var(f_i|g)$, $\sigma^2_{u,g} = Var(u_{it}|g)$ and $\sigma_{u,gts} = Cov(u_{it}, u_{is}|g)$. Note that $\sigma_{u,gts}$ degenerates to $\sigma^2_{u,g}$ if $t = s$. Then $\sigma^2_{\varepsilon,g} = \sigma^2_{f,g} + \sigma^2_{u,g}$ and $\sigma_{\varepsilon,gts} = \sigma^2_{f,g} + \sigma_{u,gts}$ in a panel population model because $f_i$ and $u_{it}$ are often assumed uncorrelated. Therefore, the partially matched subsample causes correlation between the composite error $\varepsilon_{it}$ and $\varepsilon_{is}$ via the fixed effect $f_i$ as well as the serial correlation in $u_{it}$ if there is any.

Recall that $\mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}) = -\hat{\boldsymbol{\mu}}^{\varepsilon}$. Hence, (3.8) is the generalization of the asymptotic distribution of $\sqrt{n}\mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta})$ in (2.19) to the case of partially matched samples. The asymptotic variance $\boldsymbol{\Psi}^{\varepsilon}$ in Theorem 1 is the inverse of the optimal weighting matrix to use in the pseudo panel MD estimation where samples from different periods may be partially matched, i.e. $\boldsymbol{\Psi}^{\varepsilon}$ is the "$\mathbf{M}$ matrix" in this setup. The diagonal elements of $\boldsymbol{\Psi}^{\varepsilon}$ are exactly the same as those of the $\mathbf{M}$ matrix in the standard case of Imbens & Wooldridge (2007) where samples from different time periods are independent. In other words, $\boldsymbol{\Psi}^{\varepsilon}$ and $\mathbf{M}$ only differ on their off diagonal elements. When there is no matching, $\psi_{ts,t} = \psi_{ts,s} = 0$ and $\boldsymbol{\Psi}^{\varepsilon}$ degenerates to $\mathbf{M}$.

Let $\hat{\boldsymbol{\theta}}_{\psi}$ be the optimal MD estimator using the inverse of $\hat{\boldsymbol{\Psi}}^{\varepsilon}$ as the weighting matrix, where the estimator $\hat{\boldsymbol{\Psi}}^{\varepsilon}$ of $\boldsymbol{\Psi}^{\varepsilon}$ will be defined in (3.15) in the next subsection. By the same argument that yields $\hat{\boldsymbol{\theta}}^{opt}$, $\hat{\boldsymbol{\theta}}_{\psi}$ also has a closed-form as given in

$$\hat{\boldsymbol{\theta}}_{\psi} = \left[\hat{\boldsymbol{\mu}}^{\mathbf{x}\,\prime}\left(\hat{\boldsymbol{\Psi}}^{\varepsilon}\right)^{-1}\hat{\boldsymbol{\mu}}^{\mathbf{x}}\right]^{-1}\hat{\boldsymbol{\mu}}^{\mathbf{x}\,\prime}\left(\hat{\boldsymbol{\Psi}}^{\varepsilon}\right)^{-1}\hat{\boldsymbol{\mu}}^{y}. \tag{3.10}$$

The asymptotic variance of $\hat{\boldsymbol{\theta}}_{\psi}$ can be obtained by replacing the $\mathbf{M}$ matrix in (2.23) with $\boldsymbol{\Psi}^{\varepsilon}$, i.e.,

$$Avar(\hat{\boldsymbol{\theta}}_{\psi}) = n^{-1}\left[\boldsymbol{\mu}^{\underline{\mathbf{x}}\,\prime}(\boldsymbol{\Psi}^{\varepsilon})^{-1}\boldsymbol{\mu}^{\underline{\mathbf{x}}}\right]^{-1} \tag{3.11}$$

The asymptotic standard deviations (s.d.'s hereafter) of $\hat{\boldsymbol{\theta}}_{\psi}$ are the square roots of the diagonal elements of (3.11). As a comparison, if we ignore that the samples are partially matched and use the "wrong" weighting matrix $\mathbf{M}$ as defined in (2.20), we can still obtain a consistent estimator of $\boldsymbol{\theta}$, but it is asymptotically less efficient than $\hat{\boldsymbol{\theta}}_{\psi}$ since it ignores the correlations among partially matched samples. Denote this estimator by $\hat{\boldsymbol{\theta}}_{\mathbf{M}}$. In addition, define $\Xi = \left(\boldsymbol{\mu}^{\underline{\mathbf{x}}\prime}\mathbf{M}^{-1}\boldsymbol{\mu}^{\underline{\mathbf{x}}}\right)^{-1}$ and $\mathbf{A} = \boldsymbol{\mu}^{\underline{\mathbf{x}}\prime}\mathbf{M}^{-1}\boldsymbol{\Psi}^{\varepsilon}\mathbf{M}^{-1}\boldsymbol{\mu}^{\underline{\mathbf{x}}}$. Then,

$$Avar(\hat{\boldsymbol{\theta}}_{\mathbf{M}}) = n^{-1}\Xi\mathbf{A}\Xi, \tag{3.12}$$

which is of a "sandwich" form and thus greater than (3.11) in the matrix sense (see Kodde et al. 1990 and

Jia 2019 for formal proofs). The finite-sample performance of $\hat{\boldsymbol{\theta}}_{\psi}$ and $\hat{\boldsymbol{\theta}}_{\mathbf{M}}$ will be compared in a subsequent simulation section.

Depending on the sampling design, especially the magnitude of the matching rates between two samples from different time periods, $\boldsymbol{\Psi}_g^{\varepsilon}$ may have multiple nonzero super/sub-diagonals. To provide some examples, this paper derives the structure of $\boldsymbol{\Psi}_g^{\varepsilon}$ for yearly MORG and monthly Australia LFS, of which details are given in the online appendix.

## 3.3 Estimation

This section first develops an estimator for $\boldsymbol{\Psi}_g^{\varepsilon}$ and then defines the resulting estimators for $Avar(\hat{\boldsymbol{\theta}}_{\psi})$ and $Avar(\hat{\boldsymbol{\theta}}_{\mathbf{M}})$. Theorem 1 provides a pathway to estimate the optimal weighting matrix. We can first use some initial estimator to calculate the residuals $\check{\varepsilon}_{it}$'s. Then we can use the sample variance of $\check{\varepsilon}_{it}$ in cell $(g,t)$ to estimate $\sigma^2_{\varepsilon,g}$ and the sample covariance of $\check{\varepsilon}_{it}$ and $\check{\varepsilon}_{is}$ in the intersection of cell $(g,t)$ and cell $(g,s)$ to estimate $\sigma_{\varepsilon,gts}$. Note that $\sigma_{\varepsilon,gts}$ cannot be naively estimated by the sample covariance between the residuals in cell $(g,t)$ and those in cell $(g,s)$ because that sample covariance is not well defined. Note also that $\sigma^2_{f,g}$, $\sigma^2_{u,g}$ and $\sigma_{u,gts}$ cannot be separately estimated unless we make more assumptions about the serial correlation (e.g., $\sigma_{u,gts}=0$).

Specifically, for $t=s$, $\sigma_{\varepsilon,gts}=\sigma^2_{\varepsilon,gt}=Var(\varepsilon_{it}|g)$, notice $\varepsilon_{it}$ can be estimated by the residual $\check{u}_{it}$ obtained by plugging into (2.5) the an initial estimator of $\boldsymbol{\theta}$. A commonly used candidate for the initial estimator is the pseudo panel FE estimator $\check{\boldsymbol{\theta}}$ as defined in (2.10). Then, a consistent estimator of $\sigma^2_{\varepsilon,gt}$ is

$$\hat{\sigma}^2_{\varepsilon,gt}=n_{gt}^{-1}\sum_{i\in\mathcal{I}_{gt}}r_{ig}\left(\check{u}_{it}-\tilde{u}_{gt}\right)^2 \tag{3.13}$$

where $\tilde{u}_{gt}\equiv n_{gt}^{-1}\sum_{i=1}^{n_t}r_{itg}\check{u}_{it}$ is the sample average of $\check{u}_{it}$ within cell $(g,t)$.

To estimate $\sigma_{\varepsilon,gts}$ for $t\neq s$, notice that the overlapped subsample of any two samples are matched and the matching is known. Therefore, the matched subsample can be used to estimate $\sigma_{\varepsilon,gts}$ as in

$$\hat{\sigma}_{\varepsilon,gts}=m_{gts}^{-1}\sum_{i\in\mathcal{I}_{gts}}r_{ig}\left(\check{u}_{it}-\tilde{u}_{gts}\right)\left(\check{u}_{is}-\tilde{u}_{gts}\right), for\ t\neq s, \tag{3.14}$$

where $\tilde{u}_{gts}\equiv m_{gts}^{-1}\sum_{i\in\mathcal{I}_{gts}}r_{itg}\check{u}_{it}=\tilde{u}_{gts}$. (3.13) and (3.14) together defines the estimator $\hat{\sigma}_{gts}$ for all $t$ and $s$.

Finally, recall that $\rho_g$ and $\kappa_t$ can be consistently estimated by $\hat{\rho}_{gt}=n_{gt}/n_t$ and $\hat{\kappa}_t=n_t/n$, respectively. The matching rates $\psi_{ts,t}$ and $\psi_{ts,s}$ are usually known from the sampling design, but they can also be estimated by $\hat{\psi}_{ts,t}=m_{ts}/n_t=m_{gts}/n_{gt}$ and $\hat{\psi}_{ts,s}=m_{ts}/n_s=m_{gts}/n_{gs}$, respectively, to reflect the uncertainty caused by unexpected events in sampling. Hence, the diagonal elements of $\boldsymbol{\Psi}_g^{\varepsilon}$, $\left(\rho_g\kappa_t\right)^{-1}\sigma_{\varepsilon,gts}$, can be consistently estimated by $\frac{n}{n_{gt}}\hat{\sigma}^2_{\varepsilon,gt}$ which is the same as how we estimate the diagonal elements of $\mathbf{M}_g$, and the off-diagonal elements of $\mathbf{\Psi}^{\varepsilon}_g$, $(\psi_{ts,t}\psi_{ts,s})^{\frac{1}{2}}\left(\rho_g\kappa_t\rho_g\kappa_s\right)^{-\frac{1}{2}}\sigma_{\varepsilon,gts}$, can be consistently estimated by $\frac{m_{gts}n}{n_{gt}n_{gs}}\hat{\sigma}_{\varepsilon,gts}$. The resulting consistent estimator $\hat{\mathbf{\Psi}}^{\varepsilon}_g$ contains $\frac{n}{n_{gt}}\hat{\sigma}^2_{\varepsilon,gt}$'s on the t-th principal diagonal and $\frac{m_{gts}n}{n_{gt}n_{gs}}\hat{\sigma}_{\varepsilon,gts}$ on position $(s,t)$ for $t \neq s$, and the estimator of $\mathbf{\Psi}^{\varepsilon}$ is

$$\hat{\mathbf{\Psi}}^{\varepsilon} = \mathbf{diag}\left\{\hat{\mathbf{\Psi}}^{\varepsilon}_g\right\}. \tag{3.15}$$

Correspondingly, (3.11) can be estimated by

$$\widehat{Avar(\hat{\boldsymbol{\theta}}_{\psi})} = n^{-1}\left[\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}\prime}(\hat{\mathbf{\Psi}}^{\varepsilon})^{-1}\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}}\right]^{-1} \tag{3.16}$$

and (3.12) can be estimated by

$$\widehat{Avar(\hat{\boldsymbol{\theta}}_{\mathbf{M}})} = n^{-1}\hat{\Xi}\hat{\mathbf{A}}\hat{\Xi} \tag{3.17}$$

where $\hat{\Xi} \equiv \left(\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}\prime}\mathbf{M}^{-1}\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}}\right)^{-1}$ and $\hat{\mathbf{A}} \equiv \hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}\prime}\hat{\mathbf{M}}^{-1}\hat{\mathbf{\Psi}}^{\varepsilon}\hat{\mathbf{M}}^{-1}\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}}$.

# 4 Simulation

This section briefly presents a series of carefully designed simulation cases to show the finite sample property the pseudo panel MD estimators using the proposed weighting matrix $\mathbf{\Psi}^{\varepsilon}$. The model used to simulate data is:

$$y_{it} = \beta_1 + \beta_2 x_{it} + \eta_t + \underbrace{(\alpha_{g_i} + e_i)}_{f_i} + u_{it},\ i \in \mathcal{I}_t,\ t = 1, \cdots, T, \tag{4.1}$$

where $x_{it} \sim N(gt/6, 1)$, $\beta_1 = \beta_2 = 1$, $\eta_t = t - 1$ and $\alpha_g = g - 1$. $x_{it}$ is independent of $g_i$, $e_i$ and $u_{it}$. The simulation cases to be considered differ in their variance specifications on $f_i|g$ (equivalently, $e_i|g$) and variance and serial covariance specifications on $u_{it}|g$. The yearly MORG sampling design is adopted with constant sample sizes ($n_t = n_0$ and thus $\kappa_t = 1/T$) and fixed group proportions ($\rho_{gt} = \rho_g = 1/G$). This implies a constant matching rate $\psi_{ts,t} = \psi_{ts,s} = \psi = 50\%$. Throughout the simulation, $G = 8$ and $T = 10$ is chosen as the focus specification.

The main quantities of interest in the simulation study are two MD estimators of $\beta_2$ and, more importantly, their corresponding s.e.'s. From now on, the MORG estimator and s.e. refer to those obtained from the MD estimation using $(\hat{\mathbf{\Psi}}^{\varepsilon})^{-1}$ as the weighting matrix (see (3.10) and (3.16)). The SWH (abbreviated from sandwich) estimator and s.e., on the other hand, refer to those obtained from using $(\hat{\mathbf{M}})^{-1}$ (see (2.22) and (3.17)). In addition, the so-called Naive s.e. (see (2.24)) for the SWH estimator is also studied as a comparison to the SWH s.e. The MORG and SWH s.e.'s should be consistent for their respective asymptotic counterparts that are referred to as the MORG and SWH asymptotic s.d.'s hereafter. The Naive s.e.'s, however, are generally inconsistent for the SWH asymptotic s.d.'s under partially matched samples.

To induce nonzero $\sigma_{\varepsilon,gts}$ for $s \neq t$, varying $\sigma^2_{f,g}$ over $g$ is used in relevant cases. The feature that $\sigma^2_{f,g}$ varies with $g$ is referred to as cohort-wise heteroskedasticity in $f_i$ hereafter. Besides $f_i$, another source for nonzero $\sigma_{\varepsilon,gts}$ is heteroskedasticity over group-time cells and/or serial correlation in $u_{it}$. To introduce serial correlation in $u_{it}$, the first order autoregressive model (AR(1)),

$$u_{it} = \gamma_0 u_{it-1} + \xi_{it} \tag{4.2}$$

with the initial condition $u_{i0} = 0$, is used in relevant cases. The innovation terms $\xi_{it}$'s are independent across $i$ and $t$ and follow the distribution $\xi_{it}|g \sim N(0, \sigma^2_{\xi,gt})$ where $\sigma^2_{\xi,gt} \equiv Var(\xi_{it}|g)$ could depend on $g$ and $t$. $\sigma^2_{u,gt}$ and $\sigma_{u,gts} \equiv Cov(u_{it}, u_{is}|g)$ vary with $(g,t)$ accordingly. Such varying $\sigma^2_{u,gt}$ is termed as cell-wise heteroskedasticity in $u_{it}$ hereafter, as it may depend on both $g$ and $t$.

To discover what DGP features favor the MORG estimator and s.e., four benchmark cases (Cases 1.1 to 1.4) and three extended cases (Cases 2 to 4) are designed. In the four benchmark cases, $f_i$ is homoskedastic and $u_{it}$ is homoskedastic and serially uncorrelated. These cases differ, however, on the relative magnitudes of $\sigma^2_{f,g}$ and $\sigma^2_{u,gt}$. In Case 1.1, $\sigma^2_{f,g} = 1$, $\sigma^2_{u,gt} = 100$; in Case 1.2, $\sigma^2_{f,g} = 100$, $\sigma^2_{u,gt} = 1$; in Case 1.3, $\sigma^2_{f,g} = 1$, $\sigma^2_{u,gt} = 1$; in Case 1.4, $\sigma^2_{f,g} = 100$, $\sigma^2_{u,gt} = 100$. Although block diagonal in all the four benchmark cases, $\mathbf{\Psi}^{\varepsilon}$ in Case 1.1 is numerically close to the identity matrix. In Case 1.2 where $\sigma^2_{f,g}$ dominates $\sigma^2_{u,gt}$, a numerically nontrivial block diagonal structure of $\mathbf{\Psi}^{\varepsilon}$ emerges, but the elements on each of the principal, super and sub diagonals are the same across the diagonal. Cases 1.3 and 1.4 produce essentially the same block diagonal structure of $\mathbf{\Psi}^{\varepsilon}$ that only differ by a scaling factor, and the deviation of $\mathbf{\Psi}^{\varepsilon}$ from the identity matrix is less pronounced than in Case 1.2. In fact, knowing the DGP allows us to calculate the theoretical relative differences between the asymptotic SWH and MORG s.d.'s for $\beta_2$. In Cases 1.1 to 1.4, they are are $4.55 \times 10^{-6}$, 5.84%, 0.969% and 0.969%, respectively.

Case 2 introduces the following cohort-wise heteroskedasticity in $f_i$,

$$\sigma^2_{f,g} = 100 \cdot 1_{\{g \leq G/2\}} + 1_{\{g > G/2\}}, \tag{4.3}$$

where $\sigma^2_{f,g}$ dominates $\sigma^2_{u,gt}$ for $g \leq G/2$ but is comparable to $\sigma^2_{u,gt}$ for $g > G/2$. The rest setup is the same as in Case 1.1. In Case 2, the cohort-wise heteroskedasticity in $f_i$ supposedly creates a slightly smaller difference between $\mathbf{\Psi}^{\varepsilon}$ and $\mathbf{M}$ than in Case 1.2. Consequently, a smaller theoretical relative difference between the asymptotic SWH and MORG s.d.'s, 2.19%, appears.

In Case 3, $f_i$ reverts to the homoskedastic specification with $\sigma^2_{f,g} = 1$, but $u_{it}$ is cell-wise heteroskedastic and serial correlated. Specifically, $u_{it}$ follows (4.2) with $\gamma_0 = -0.95$, and $\sigma^2_{\xi,gt}$ follows the specification

$$\sigma^2_{\xi,gt} = Var(\xi_{it}|g) = \max\left\{1, \left[\sigma^2_{b,gt} - \gamma_0^2\sigma^2_{b,g(t-1)}\right]\right\}. \tag{4.4}$$

where, for real numbers $a$ and $p$, $\sigma^2_{b,gt} \equiv b_{gt}\left[\sin\left(a\frac{gt}{GT}\right)\right]^p$ and $b_{gt} \equiv 10 \cdot 1_A + 100 \cdot 1_{\bar{A}}$ with $A = \{g \leq G/2 \; or \; t \leq T/2\}$. (4.4) essentially endows $\sigma^2_{u,gt}$ with the pattern of $\sigma^2_{b,gt}$ and at the same time ensures that $\sigma^2_{\xi,gt} > 0$. The parameter values used is $(\gamma_0, a, p) = (-0.95, 3.1415, 0.5)$. The resulting deviation of $\mathbf{\Psi}^\varepsilon$ from $\mathbf{M}$ is prominent: The theoritical relative difference between the asymptotic SWH and MORG s.d.'s for $\beta_2$ is 16.80%.

The ultimate case, Case 4, merges the feature of cohort-wise heteroskedastic $f_i$ in Case 2 and that of the cell-wise heteroskedastic and serial correlated $u_{it}$ in Case 3, creating the wildest variance/covariance variation in the composite error in the simulation study. The resulting relative difference in the asymptotic SWH and MORG s.d.'s. is 22.56%.

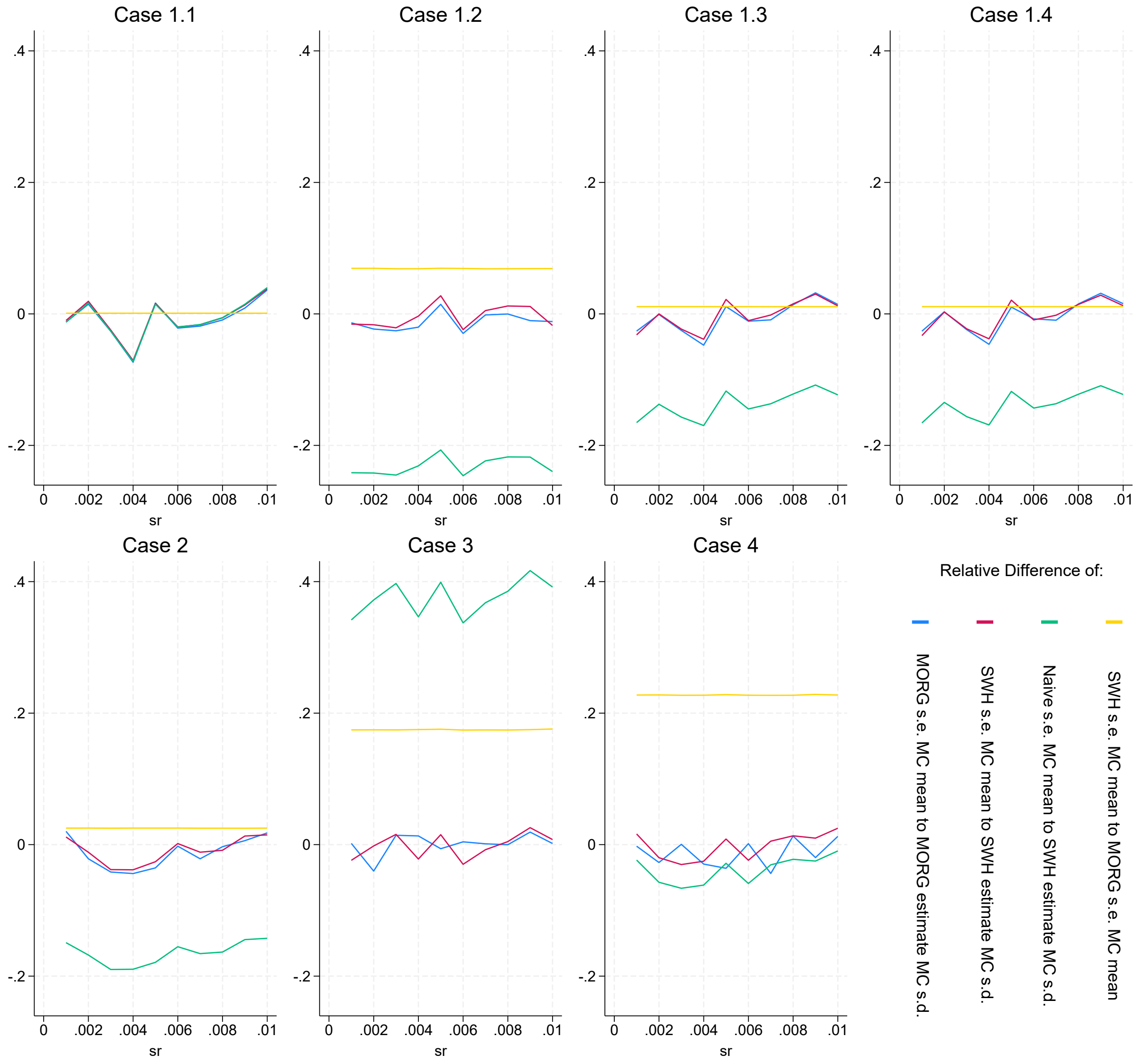
To mimic the fact that the population is finite in reality, finite populations are used in the simulation. Moreover, varying sampling rates are considered as a robustness check. For each simulation case defined above, the sampling rate varies from 0.1% to 1% with a step size of 0.1%, whereas the finite population cohort size shrinks from $240,000$ to $24,000$ accordingly so that the realized sample cohort size is approximately 240. This design eliminates the effect of varying sample size on inference. Rigorously speaking, using finite populations requires some adjustment to inference, but the adjustment is negligible here given the large population sizes and small sampling rates used.[10]

Figure 4.1 presents a summary of the key simulation findings based on 1000 replications. More detailed results can be found in the online appendix. Two observations stand out. First, the relative differences of the MORG and SWH s.e.'s w.r.t to their corresponding Monte Carlo (MC) s.d.'s (blue and red lines) fluctuate around zero in all 7 cases, which are finite-sample evidence for the consistency of the MORG and SWH s.e.'s. That is not the case for the Naive s.e. (green lines) when $\mathbf{\Psi}^\varepsilon$ differs from the identity matrix (all cases except 1.1). Therefore, ignoring dependence among samples and naively use $\mathbf{M}$ as the optimal weighting matrix generally yields biased inference. Note that the finite-sample bias in the Naive s.e. in Case 4 is not as noticeable as that in Case 2 or 3. This is probably attributable to the finite-sample biases in the Naive s.e.'s in Cases 2 and 3 having opposite signs, which more or less offset each other when aggregated in Case 4.

Secondly, and perhaps more importantly, when $\mathbf{\Psi}^\varepsilon$ differs non-trivially from $\mathbf{M}$ (Cases 1.2, 2, 3 and 4), we get nontrivial efficiency gains in MORG s.e.s over SWH s.e.s (yellow lines). This is as expected. Also, the

[10] In fact, by Li & Ding 2017, using finite populations merits a multiplicative shrinkage factor $\sqrt{1-\lambda}$ on the s.e.'s, where $\lambda$ is the sampling rate. For $\lambda$ in $(0.1\%, 1\%)$, $\sqrt{1-\lambda}$ lies in $(0.9950, 0.9995)$.

Figure 4.1: Cases 1.1 to 4 simulation results for $G = 6$ and $T = 10$ from 1000 replications. The sampling rate (sr) varies from 0.1% to 1% with step 0.1%. The finite population size varies accordingly so that the sample cohort size in each period is roughly 240.



observed efficiency gains well match their theoretical values: Averaging the results over the 10 sampling rates, the SWH s.e.'s are .1038%, 6.899%, 1.097% and 1.101% greater than the MORG s.e.'s for Cases 1.1 to 1.4, respectively; for Cases 2 to 4, the differences are 2.19%, 16.80% and 17.48%, respectively. In addition, Cases 1.2, 2, 3 and 4 highlight that the relative magnitude of the group-time cell variance of the fixed effect w.r.t. that of the idiosyncratic error, cohort-wise heteroskedasticity in the fixed effect, cell-wise heteroskedasticity and/or serial correlation in the idiosyncratic error are typical sources for potential efficiency gains of using $\mathbf{\Psi}^{\varepsilon}$.

## 5 An Empirical Illustration

As an illustration, this section applies the MORG and SWH estimators and their s.e.'s to the classical empirical question of estimating monetary returns to education.[11] This analysis uses the MORG files spanning from 2010 to 2019 ($T = 10$). The specification is similar to that in Angrist & Krueger (1991). Specifically, the sample are restricted to black and white men only. The dependent variable used is the logarithm of hourly earnings. The key regressor is education measured in years. A race dummy, a marital status dummy and a metropolitan status dummy (standard metropolitan statistical area, or SMSA) are included in the control variable list. The marital status dummy is defined so that married civilian spouse present, married armed force spouse present and married spouse absent or separated are grouped in to the married group whereas widowed, divorced, separated and never married are grouped into the other group. Age and age squared are also included to capture the potential nonlinear age effects. In addition, 9 region dummies are included to control spatial variations. The samples are further restricted to individuals aged 26-55 in 2018 (born 1963-1992) and are divided to six 5-year birth cohorts ($G = 6$).

The main results are presented in Table 1. To save space, the estimates on the the region dummies are omitted, and additional estimates separately obtained from 2010-2014 and 2015-2019 are provided in the online appendix. The results from pooled OLS (POLS, first two columns) are reported for comparison. For the sake of brevity, the discussion focuses solely on the results on education. Consistent with the theory and the simulation results discussed earlier, the MORG s.e. for education is smaller than the SWH s.e. no matter the age function is included or not in the specification (Column 3 vs 5, Column 4 v.s. 6). Specifically, the SWH s.e. is approximately 14% greater than the MORG s.e in the absence of the age function, and 11% greater when the age function is included. Furthermore, there is a noticeable discrepancy between the Naive s.e. and the SWH s.e. under both specifications, indicating potential bias in the Naive s.e. for this particular application.

[11] For surveys of this literature, see Card 1999, Card 2001, Heckman et al. 2006, McMahon 2009 and Oreopoulos & Petronijevic 2013 among others.

Table 1: CPS, 2010-2019

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| | PPOLS | POLS | MORG | MORG | SWH | SWH | Naive | Naive |
| Years of Edu. | 0.085$^{***}$ | 0.084$^{***}$ | 0.249$^{***}$ | 0.218$^{***}$ | 0.294$^{***}$ | 0.245$^{***}$ | 0.294$^{***}$ | 0.245$^{***}$ |
| | (0.000) | (0.000) | (0.053) | (0.052) | (0.063) | (0.061) | (0.068) | (0.067) |
| Married | 0.141$^{***}$ | 0.127$^{***}$ | 0.675$^{***}$ | 0.297 | 0.688$^{***}$ | 0.319 | 0.688$^{***}$ | 0.319 |
| | (0.002) | (0.002) | (0.120) | (0.164) | (0.131) | (0.183) | (0.121) | (0.165) |
| Black | -0.184$^{***}$ | -0.187$^{***}$ | 0.306 | 0.009 | 1.109 | 0.498 | 1.109 | 0.498 |
| | (0.003) | (0.003) | (0.801) | (0.681) | (0.922) | (0.781) | (0.897) | (0.768) |
| Age | | 0.083$^{***}$ | | 0.008 | | -0.035 | | -0.035 |
| | | (0.001) | | (0.099) | | (0.117) | | (0.121) |
| Age Squared | | -0.001$^{***}$ | | -0.000$^{**}$ | | -0.000$^{*}$ | | -0.000$^{**}$ |
| | | (0.000) | | (0.000) | | (0.000) | | (0.000) |
| Metropolitan | 0.081$^{***}$ | 0.081$^{***}$ | 0.224 | 0.344 | 0.192 | 0.365 | 0.192 | 0.365 |
| | (0.002) | (0.002) | (0.544) | (0.468) | (0.703) | (0.586) | (0.803) | (0.671) |
| R-squared | 0.269 | 0.276 | | | | | | |
| N | 449568 | 449568 | | | | | | |

Years 2010-2019

Results on region dummies, group dummies and time dummies are omitted.

$^{*}$ $p < 0.05$, $^{**}$ $p < 0.01$, $^{***}$ $p < 0.001$

Although the primary focus of the paper is not on the MD coefficient estimates, their deviation from the conventional return-to-education estimates in the literature warrants a discussion. In the presence of the age function, the estimated monetary returns to one more year of education in Table 1 are 22% and 23% for the MORG and SWH estimators, respectively, and are statistically significant. These estimates are approximately double the usual estimates of around 10% found in the literature. In comparison, the POLS estimator, regardless of whether the age function is used, yields a more stable and lower return around 8%, which is in the ballpark of the literature. The inclusion of the age function significantly impacts the MD estimates but has minimal effect on the POLS estimates.

The substantial discrepancy between POLS and MD can be attributed to the different variations they use to estimate the coefficients in the underlying linear panel date model. The MD estimators essentially are weighted cohort-level FE estimators – they use only the cohort-level within variation over time for each birth cohort to estimate the return to education. POLS, on the other hand, explores both the cohort-level within variation and the within-cell variation at the individual level. The within-cell variation may reveal a very different association between wage and education from the that embedded in the cohort-level within variation, as there may exist individual-level specific correlations between the idiosyncratic error ($u_{it}$) and education.[12] When projected on to the cohort level, this individual-level correlation is only be partially accounted for if the group membership is not exogenous w.r.t. the idiosyncratic error. There will be remaining correlation between the cell mean of the idiosyncratic error, $E(u_{it}|g_i)$, and that of education, $E(edu_{it}|g_i)$, which would

[12] Similar for association between wage and other covariates. Dropped hereafter to simply the interpretation.

bias the MD estimators.

Is the group membership exogenous under the current specification used? Probably not. For example, Card & Lemieux (2001) show that falling relative supply of educated workers is one important factor in explaining the widening of the college-high school wage gap during the 1980's and 1990's. This factor, however, is not controlled in the analysis here.[13] Hence, the MD estimators may still be biased, although they accounted for the fixed effect and thus removed the bias induced by the fixed effect. The POLS estimators are also biased because they fail to address the individual-level fixed effect, and education is likely endogenous as well. In general, POLS and MD will be close to each other only if the are both unbiased, i.e., if education is exogenous to both the fixed effect and the idiosyncratic error at the individual level and the group membership is exogenous to the idiosyncratic error.

## 6 Concluding Remarks

The partially-matched-sample correction to the optimal weighting matrix in MD pseudo panel estimation is motivated by the observation that dependence between samples from different periods may arise due to partially matched sampling designs such as CPS in the U.S. and LFS in Australia. In this paper, using this correction under partially matched sampling has been shown to be effective in achieving significant efficiency gains in both a simulation study and an empirical application. Future extensions can consider generalizations to dynamic models and unequally spaced pseudo panels; a closer look at the different variations explored for identification by approaches on individual-level data (such as POLS) and those on cohort-level data (such as pseudo panel MD) would also benefit the literature.

## References

Angrist, J. D. & Krueger, A. B. (1991), 'Does Compulsory School Attendance Affect Schooling and Earnings?', *The Quarterly Journal of Economics* **106**(4), 979–1014.
**URL:** *http://www.jstor.org/stable/2937954*

Antman, F. & McKenzie, D. J. (2007), 'Earnings mobility and measurement error: A pseudo-panel approach', *Economic Development and Cultural Change* **56**(1), 125–161.

Blundell, R., Meghir, C. & Neves, P. (1993), 'Labour supply and intertemporal substitution', *Journal of Econometrics* **59**(1-2), 137–160.

[13] Card & Lemieux (2001) use the Census data sets to construct relative supply measures, which is not directly replicable with the MORG files. Since finding a causal return to education is not the focus here, this paper adheres to the specification used by Angrist & Krueger (1991).

Browning, M., Deaton, A. & Irish, M. (1985), 'A profitable approach to labor supply and commodity demands over the life-cycle', *Econometrica: journal of the econometric society* pp. 503–543.

Campbell, D. L. & Lusher, L. (2019), 'The impact of real exchange rate shocks on manufacturing workers: An autopsy from the MORG', *Journal of International Money and Finance* **91**, 12–28.

Card, D. (1999), 'The causal effect of education on earnings', *Handbook of labor economics* **3**, 1801–1863.

Card, D. (2001), 'Estimating the return to schooling: Progress on some persistent econometric problems', *Econometrica* **69**(5), 1127–1160.

Card, D. & Lemieux, T. (2001), 'Can falling supply explain the rising return to college for younger men? A cohort-based analysis', *The quarterly journal of economics* **116**(2), 705–746.

Collado, M. D. (1997), 'Estimating dynamic models from time series of independent cross-sections', *Journal of Econometrics* **82**(1), 37–62.

CPS Technical Documentation (2014), Redesign of the Sample for the Current Population Survey, Technical report, Bureau of Labor Statistics.
**URL:** *https://www.bls.gov/cps/sample_redesign_2014.pdf*

Dang, H.-A. H. & Lanjouw, P. F. (2023), 'Measuring poverty dynamics with synthetic panels based on repeated cross sections', *Oxford Bulletin of Economics and Statistics* **85**(3), 599–622.

Dang, H.-A., Lanjouw, P., Luoto, J. & McKenzie, D. (2014), 'Using repeated cross-sections to explore movements into and out of poverty', *Journal of Development Economics* **107**, 112–128.
**URL:** *http://www.sciencedirect.com/science/article/pii/S0304387813001521*

Deaton, A. (1985), 'Panel data from time series of cross-sections', *Journal of econometrics* **30**(1), 109–126.

Gardes, F., Duncan, G. J., Gaubert, P., Gurgand, M. & Starzec, C. (2005), 'Panel and pseudo-panel estimation of cross-sectional and time series elasticities of food consumption: The case of us and polish data', *Journal of Business & Economic Statistics* **23**(2), 242–253.

Heckman, J. J., Lochner, L. J. & Todd, P. E. (2006), 'Earnings functions, rates of return and treatment effects: The Mincer equation and beyond', *Handbook of the Economics of Education* **1**, 307–458.

Imbens, G. W. & Wooldridge, J. M. (2007), *What's new in econometrics?*, NBER.

Inoue, A. (2008), 'Efficient estimation and inference in linear pseudo-panel data models', *Journal of Econometrics* **142**(1), 449–466.

Jia, F. (2019), Redundancy of Additional Restrictions in Minimum Distance Estimation.

Jones, S., Sohnesen, T. P. & Trifkovic, N. (2023), 'Educational expansion and shifting private returns to education: Evidence from Mozambique', *Journal of International Development* .

Juodis, A. (2018), 'Pseudo Panel Data Models With Cohort Interactive Effects', *Journal of Business and Economic Statistics* **36**(1), 47–61.

Kodde, D. A., Plam, F. C. & Pfann, G. A. (1990), 'Asymptotic least-squares estimation efficiency considerations and applications', *Journal of Applied Econometrics* **5**(3), 229–243.

Li, X. & Ding, P. (2017), 'General forms of finite population central limit theorems with applications to causal inference', *Journal of the American Statistical Association* **112**(520), 1759–1769.

McKenzie, D. J. (2001), 'Estimation of AR (1) models with unequally spaced pseudo-panels', *The Econometrics Journal* **4**(1), 89–108.

McKenzie, D. J. (2004), 'Asymptotic theory for heterogeneous dynamic pseudo-panels', *Journal of Econometrics* **120**(2), 235–262.

McMahon, W. W. (2009), *Higher learning, greater good: The private and social benefits of higher education*, JHU Press.

Meng, Y., Brennan, A., Purshouse, R., Hill-McManus, D., Angus, C., Holmes, J. & Meier, P. S. (2014), 'Estimation of own and cross price elasticities of alcohol demand in the UK - A pseudo-panel approach using the Living Costs and Food Survey 2001–2009', *Journal of health economics* **34**, 96–103.

Moffitt, R. (1993), 'Identification and estimation of dynamic models with a time series of repeated cross-sections', *Journal of Econometrics* **59**(1), 99–123.

Moretti, E. (2004), 'Estimating the social return to higher education: evidence from longitudinal and repeated cross-sectional data', *Journal of econometrics* **121**(1), 175–212.

Newey, W. K. & McFadden, D. (1994), 'Large sample estimation and hypothesis testing', *Handbook of econometrics* **4**, 2111–2245.

Oreopoulos, P. & Petronijevic, U. (2013), 'Making college worth it: A review of research on the returns to higher education'.

Saksena, M. & Maldonado, N. (2017), 'A dynamic estimation of obesity using Nhanes data: a pseudo-panel approach', *Health Economics* **26**(12), e140—-e159.

Verbeek, M. (2008), Pseudo-panels and repeated cross-sections, *in* 'The Econometrics of Panel Data', Springer, pp. 369–383.

Verbeek, M. & Nijman, T. (1993), 'Minimum MSE estimation of a regression model with fixed effects from a series of cross-sections', *Journal of Econometrics* **59**(1), 125–136.

Verbeek, M. & Vella, F. (2005), 'Estimating dynamic models from repeated cross-sections', *Journal of econometrics* **127**(1), 83–102.

Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data*, 2nd edn, MIT Press, Boston MA.